

This is a repository copy of *Carbohydrate anomalies in the PDB*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/95242/>

Version: Accepted Version

Article:

Agirre, Jon orcid.org/0000-0002-1086-0253, Davies, Gideon orcid.org/0000-0002-7343-776X, Wilson, Keith orcid.org/0000-0002-3581-2194 et al. (1 more author) (2015) Carbohydrate anomalies in the PDB. NATURE CHEMICAL BIOLOGY. p. 303. ISSN 1552-4450

<https://doi.org/10.1038/nchembio.1798>

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Carbohydrate anomalies in the PDB

Jon Agirre, Gideon Davies, Keith Wilson & Kevin Cowtan

York Structural Biology Laboratory, Department of Chemistry, The University of York, England.

Nature Chemical Biology **11**, 303 (2015) doi:10.1038/nchembio.1798

Published online 17 April 2015

Erratum (July, 2015)

The importance of carbohydrates both to fundamental cellular biology and as integral parts of therapeutics (including antibodies) continues to grow. Correct glycans are important for FDA drug approval, and statistical analyses of glycan structures reported in the Protein Data Bank (PDB) are increasingly used by diverse communities. However, carbohydrates (and other small molecules) are often handled poorly in macromolecular structural biology. When such small molecules are present in macromolecule structures, they are often reported with stereo and region-chemical errors and in unlikely conformations. Stereo and regio-chemistry should always be correct and whilst conformational distortions may reflect interactions taking place in a complex¹, most are also likely to be erroneous - resulting from poor chemical understanding and lack of appropriate stereochemical restraints in refinement, often against low-resolution data².

Pyranoside sugars have clear conformational preferences dictated by a minimization of angle, torsional and trans-annular strains, resulting in a favored chair (C) conformation for all biologically-relevant pyranoses. While carbohydrate-active enzymes can force a distortion from this minimal-energy conformation to a higher-energy one in order to enable catalysis³, this is a specialized event; of all the sugars deposited in the PDB, 65% sit undisturbed on N-glycosylation trees.

In the course of analyzing the coordinates of monosaccharides in the PDB we unveiled a problem with regard to carbohydrate conformations. Using the Privateer⁴ software, we computed the real space correlation coefficient (RSCC) against positive omit electron density (as a bias-minimized fit to observed density) for all N-glycan-forming D-pyranosides from the PDB as well as calculating their conformation-determining Cremer-Pople puckering parameters⁵. This subgroup was chosen as all are expected to be in the preferred ⁴C₁ chair conformation.

A number of anomalies appear from this analysis. Firstly, 64% of all N-glycan D-pyranosides have RSCCs < 0.8 reflecting poor fit to the electron density. Indeed 12% have RSCC < 0.5. Reflecting comments by others on the problems of carbohydrate nomenclature in the PDB⁶, it is clear that a number of models have been built incorrectly from the start – approximately 5% of the N-acetyl-D-glucosamine (GlcNAc) moieties attached to asparagine residues show an incorrect α -linkage, and some monosaccharides are then built "upside-down", resulting in a ¹C₄ conformation. In addition, there are more subtle errors reflecting incorrect refinement

of the deposited sugar that can nevertheless have enormous implications when interpreted in a biological context.

A plot of ring conformation against resolution of the X-ray analysis, colored by RSCC, is shown in Figure 1. As expected, at atomic resolutions ($< 1.2\text{\AA}$, model precision better than 0.01\AA), all sugars showing high RSCC are 4C_1 chairs, corresponding to the yellow cluster at the top left of the figure. However, as the resolution gets lower and model precision poorer, unexpected higher-energy conformations start to appear. Most of these models also show low RSCC (< 0.8), seen as blue entries spread across the other conformations in the figure.

Whilst energetically unfavorable models may reflect a poor knowledge of glycochemistry and "optimistic" density interpretation (reflected in low RSCCs), it is nevertheless clear that macromolecular crystallographers are failing to apply appropriate conformational restraints to encourage chemically sensible models at lower resolutions ($> 1.6\text{\AA}$). Although community re-refinement efforts such as PDB_REDO⁷ have led to substantial improvement in protein models, many sugars remain in high-energy conformations due to re-refinement without dihedral restraints. Torsion restraints, which approximate the eponymous energy barriers, can be used to penalize models with eclipsed conformations, encouraging a particular ring puckering for sugars. However, the perceived difficulty in modeling torsional preferences often results in these restraints being tacitly turned-off, regardless of the resolution, in many refinement and model-building programs. This creates a vicious circle: publication and deposition of incorrect structures informs subsequent statistical analyses that lead the community to believe that the deposited structures are "normal". This circle must be broken to prevent continued deposition of conformations that are chemically unlikely.

Problems with the refinement of protein structures in the 1980's led to the rise of standard dictionaries⁸, consistent refinement strategies, better graphics programs and community-accepted best practice (cross validation, deposition of coordinates and data, etc.). Ligands generally, and carbohydrates especially, got left behind. The fundamental roles of carbohydrates in cell biology and medicine, the extraordinary experimental advances in carbohydrate synthesis and the large increase in glycosylation-competent eukaryotic expression systems now demand improved refinement protocols for these key biological species too.

Jon Agirre, Gideon Davies, Keith Wilson & Kevin Cowtan

York Structural Biology Laboratory, Department of Chemistry, The University of York, England.

e-mail: kevin.cowtan@york.ac.uk or jon.agirre@york.ac.uk

¹ Davies, G *et al.* Acc Chem Res, 2012, 45(2):308-16.

² Reynolds, CH. ACS Med. Chem. Lett., 2014, 5(7):727–729.

³ Davies, G *et al.* Biochem J, 1997, 321(2):557–559.

⁴ Agirre, J and Cowtan, K. Computational Crystallography Newsletter, 2015, Jan:10-12.

⁵ Cremer, D and Pople, JA, J. Am. Chem. Soc., 1975, 97 (6):1354–1358.

⁶ Lütke, T. *et al.* Carbohydr Res., 2004, 339(5):1015-20.

⁷ Joosten, RP *et al.* Bioinformatics, 2011, 27:3392-3398.

⁸ Engh, RA and Huber, R, Acta Cryst A, 1991, 47:392-400.

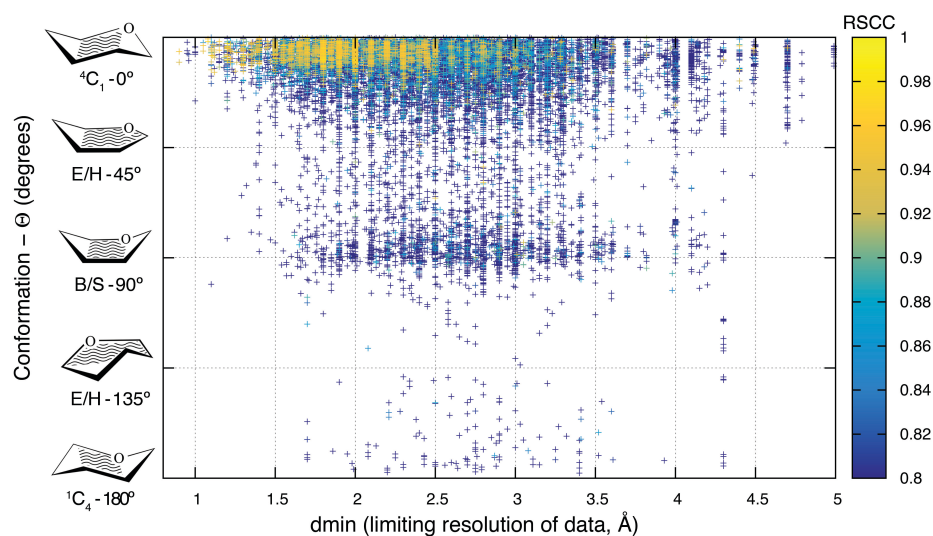


Figure 1. Distribution of D-Pyranoside ring conformations as a function of resolution for all N-linked sugars (at distance $< 2.0\text{\AA}$) in the PDB by January 2015, identified by their Chemical Component ID's: NAG, NDG, MAN, BMA, BGC, GLC, GAL and GLA. E/H: Envelopes and Half-chairs, B/S: Boats and Skew-boats, with wavy lines denoting the main ring plane. For clarity, an envelope is depicted at $\theta=45^\circ$, a half-chair at $\theta=135^\circ$, and skew-boat is omitted from the equator.